# Device effect on panoramic video+context tasks

# Device Effect on Panoramic Video+Context Tasks

Fabrizio Pece[1,4] James Tompkin[2,3] Hanspeter Pfister[3] Jan Kautz[1,5] Christian Theobalt[2]

[1]University College London [2]MPI for Informatics [3]Harvard University [4]ETH Zürich [5]NVIDIA

## ABSTRACT

Panoramic imagery is viewed daily by thousands of people, and panoramic video imagery is becoming more common. This imagery is viewed on many different devices with different properties, and the effect of these differences on spatio-temporal task performance is yet untested on these imagery. We adapt a novel panoramic video interface and conduct a user study to discover whether display type affects spatio-temporal reasoning task performance across desktop monitor, tablet, and head-mounted displays. We discover that, in our complex reasoning task, HMDs are as effective as desktop displays even if participants felt less capable, but tablets were less effective than desktop displays even though participants felt just as capable. Our results impact virtual tourism, telepresence, and surveillance applications, and so we state the design implications of our results for panoramic imagery systems.

## Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Evaluation/methodology, Interaction styles; H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities, Evaluation/methodology.

## Keywords

Panoramas; immersion; multidisplay adaption; video.

## 1. INTRODUCTION

Panoramic images and video are now common, with the world quickly being mapped at street level by companies and tourists alike. Every day, thousands of people view panoramic imagery through services like Google Streetview, viewed on different devices both exocentrically on desktop computers and egocentrically on mobile devices with orientation rotation controls. Panoramas are exploited for surveillance and collaborative telepresence applications, where spatial understanding of the scene is critical [14, 18, 22]. Furthermore, novel interfaces for exploring video collections extend spatial reasoning to include temporal reasoning [23, 37].

However, while the wide-spread use of existing panoramic imagery applications and the novelty of panoramic video applications is apparent, no research has yet investigated the possible effect of different devices on these panoramic imagery applications. This problem is brought into focus when we consider that some tasks which involve panoramic video imagery are performance critical, such as rescue services telepresence or security surveillance review. One field that has investigated the question of display effect is virtual reality, where it is suggested that immersive displays, such as head-mounted displays (HMD), may improve user performance in tasks that require a high level of spatial reasoning [20, 16, 34]. However, these works often use 3D virtual environments, and panoramas are not 3D: on a spectrum between 3D virtual environments and 2D images, panoramas are a hybrid in between — a 360° panorama can surround a user, but the scene has only spherical geometry and is effectively flat. The user cannot move from their point of view and so does not receive parallax cues. Thus, these research outcomes are not directly applicable.

The question of viewing and interacting device effect is broad, and so we begin to explore this question with a user study investigating the impact of common devices on panoramic video scene task performance. We develop a multidisplay adapting interface for viewing videos within a panoramic context. We employ this interface across three devices which sample interesting points within the design space: 1) static flat display desktop devices, which are exocentric and not immersive; 2) mobile tablet devices, which are free to rotate and act as windows into the world with orientation tracking — these are egocentric but not immersive; and 3) HMDs with orientation tracking, which are both egocentric and immersive. This range of devices covers both very common display types in desktops and tablets, and more exotic HMD displays which, with their recent affordability, are becoming more common. Furthermore, each display type forces a change of pointing interface [21], and the usability of highly spatial systems is related to their input interfaces [7, 11].

Importantly, there is no clear application boundary which limits each display type, and each instance can potentially serve a variety of applications which require panoramic video imagery: desktop displays could be used for surveillance applications and event monitoring, tablet devices could be employed for virtual tourism, while HMDs could be adopted for immersive visualization and telepresence.

Our contribution is a user study investigating device effect on the previously unexplored hybrid space of panoramic imagery. Through a simple and a complex tasks requiring the localization, recognition, and tracking of people within panoramic scenes, and through qualitative questionnaires, we find that all three displays are equally

performant for our simple task, but that for our more complex task, our tablet interface was less accurate and took more time than the HMD and desktop displays. Interestingly, participants perceived the tablet to be just as capable and usable as the desktop, even though task performance was worse. Contrarily, participants perceived the HMD to be less capable and less usable than the desktop, even though task performance was the same.

Any display adaptation then involves a trade-off, and our study has implications for this design space. Through our multidisplay adaption, and with the results of our user study, we discuss how participants responded to the different interfaces, how they approached the tasks on each of the displays, and how they evaluated the displays in usability and task-related questionnaires. We discover that exocentric views are preferred on our desktop display and that touch rotation is preferred over arm rotation on our tablet display. We combine this information to try to generalize our specific findings, and so formulate implications for designing panoramic imagery systems to aid future research and development.

# 2. RELATED WORK

Investigating device types for panoramic imagery concerns both literature on the effect of different displays across a variety of tasks, and on the scope of panoramic imagery techniques within which to try to measure an effect.

## 2.1 Display Effects

Researchers have tried to quantify the effect of display devices on user performance. Several experiments [2, 36] have compared immersive displays, such as CAVEs or HMDs, to traditional displays. The early work of Slater focuses on how immersive displays might afford users a greater sense of presence [31, 30, 28], and his studies discover that immersion can lead to increased performance in 3D spatial tasks. Bowman et al. [2] investigated human behavior and performance when using an HMD and a CAVE, discovering that HMD users are significantly more likely than CAVE users to use natural rotation in a VE. Other research measures the relationship between display type and spatial reasoning [21, 26]. They find that, along with HMDs, large projection screen systems can also offer a greater sense of immersion which may lead to better performance. Mizell et al. find that immersive displays can better convey the sense of space than desktop displays [16]. Patrick et al. [20] compare various displays which occupy comparable visual angles, and find that, while users performed significantly worse in forming cognitive maps on a desktop monitor, users performed no differently using a head-mounted display or a large projection display. Similarly, Tan et al. [34] studied the effect of large projected wall displays, and suggest that large displays afford a greater sense of presence, leading to better performance.

All the studies presented so far focus on comparing different display types while keeping the rendered content unmodified. Polys et al. [24] reverse this approach and investigate the effect of software field of view (SFOV) on user performance. The authors find that, for similar displays, higher SFOVs benefit search tasks by showing more of a scene in the periphery, but worsen accuracy in the comparison task by distorting a scene object's spatial location.

Mobile devices can also provide an egocentric view. Previously, this might have been accomplished with a boom chameleon [38], but modern systems use orientation tracking derived from embedded MEMS sensors. Ozbek et al. [19] compared video-see-through (VST) vs. optical-see-through (OST) displays for AR applications, concluding that OST devices are more suitable. Wither et al. [41] evaluate selection and annotation tasks between an HMD, a mobile VST display, and a mobile non-VST display, and found that mobile VST displays are fastest for cursor movement. Braun et al. [3] compared four different mobile AR interfaces, and found that tablet-style devices (Ultra Mobile PCs in the study) are preferred to HMDs for certain AR applications. All of these works assume in-situ browsing, and so do not compare desktop devices to mobile devices for panoramic imagery. Egocentric views on tablet interfaces require careful consideration of inputs [39].

## 2.2 Spatiality and Panoramic Imagery

The concept of spatiality was firstly introduced by Benford et al. [1], and subsequently further explored by Jensen [9, 8]. Spatiality refers to the ability of a system to support fundamental physical spatial properties such as distance, orientation, movement, and a shared frame of reference. Systems that aim to support virtual exploration of remote locations can be characterized by their degree of spatiality, with the least spatial systems supporting only the fundamental spatial property of containment, and the most spatial system facilitating user spatial thinking and immersion.

Panoramas increase spatiality with wide or omni-directional views of an environment in a single image, and so are often used for virtual environment navigation. Lippman pioneered this field with the *Movie-maps* hypermedia system [13], and his team were the first to enable "virtual travels" using omni-directional photographs geolocated to maps. More than a decade later, Chen presented *QuickTime VR* [4], where multiple perspective photographs are aligned to create 360° cylindrical panoramas for virtual space exploration.

Some works have studied the effect of panoramas on immersion [14, 35], concluding that the effort to create panoramic video might be warranted when high presence is desired [5]. However, panoramas, particularly omni-directional imagery, can be rendered in a variety of ways such as with perspective or equirectangular projections, and some spatial properties, such as distance and orientation, are warped with equirectangular projection. Recent work explores the influence of projection on how users are able to locate scene objects, concluding that 360° equirectangular visualization of the panorama is more important for whole scene object localization than maintaining real-world image features such as straight lines [17].

While significant spatial perception differences have been demonstrated between real world and virtual environments (VE) [42, 25], Willemsen and Gooch [40] additionally demonstrate that users perform equally when judging distances in a panorama-based VE and a graphics-based VE when viewed through an HMD, but that performance is worse than when viewing the real world: viewing through a display, and not the content depiction, is the important factor in their experiment. This suggests that panoramic content rendered in an experiment will not impact user spatial perception of distance over 3D graphics, which implies that display effect experiment results with 3D graphics may translate to panoramas.

Recently, the research community has developed new ways to display videos within panoramas. Unlike static panoramic images, these applications require both spatial and temporal reasoning about the scene, and so are good applications within which to assess performance. Pirk et al. [23] embed single or multiple video windows within panoramic imagery. Norris et al. [18] discover that focus-in-context video systems can enhance the sense of spatiality, and allow

users to point implicitly, explicitly, and to reference objects in local and remote spaces. Pece et al. [22] find that scene understanding is improved as more of the panorama gains a temporal dimension. Tompkin et al. [37] demonstrate that panoramic video representations improve spatial and temporal understanding when compared to existing video collection exploration interfaces. However, to our knowledge, nobody has investigated whether different devices can affect task performance for panoramic video imagery. Furthermore, many display types can naturally apply to panoramic imagery and, to our knowledge, nobody has looked at which display types are better for panoramic imagery tasks.

# 3. INTERFACE, TASKS, AND DISPLAYS

To compare device effect, we wish to design a quantifiable task for participants to undertake. Previous works on 3D display immersion have required complex spatial tasks like 3D Chess [30], and so to try and engage participants in hybrid panoramic space, the tasks should be more than just pointing within a panorama (c.f. [17]). We take inspiration from recent video+context applications which embed aligned video windows into panoramas [23, 22]: The 360° nature of the imagery in these applications allows us to compare different display types, and visualizing multiple video windows within the same panorama meets our need for an engaging tasks. Video windows are captured on tripods or with handheld cameras, and so each window is free to move within the panorama and capture action across the space of the panorama. Video windows are also captured at different, potentially overlapping time spans, and so we can include temporal reasoning tasks too. With these video windows, we can ask participants to follow or track objects, such as people, and so we ask participants to reason about the identity and whereabouts of people in space and time.

## 3.1 Software Interface

We develop a software interface which presents video windows within a panoramic context (Fig. 1), where the windows are aligned to the panorama using feature-matching techniques from computer vision. Video window selection occurs in a thumbnail strip panel along the top edge of the interface, which shows all possible videos at all times and can be scrolled horizontally. When a video is selected from thumbnail strip to be made visible as a window in the panorama, it starts playing from its beginning, and a local timelime appears at the bottom of the interface to allow temporal playback control, similar to existing video players. Adjusting the local timeline affects both the dynamic action within the video and the spatial position of the video within the panorama, as during capture the cameras were free to rotate. A second global timeline displays the temporal extent of all video windows at once. Adjusting the global timeline synchronously adjusts the playback of all video windows relative to global time, and allows the visualization of events which share the same time but otherwise have no visual overlap.

It is important that this interface is able to adapt with minimal changes to different display types. First, the interface allows equirectangular map projection with an infinite-pan canvas, and look-around perspective projection. This allows the interface to adapt to exocentric displays with map projection, and to egocentric displays with orientation tracking for perspective projection. Both projection modes allow zooming into the scene, which should allow participants to overcome any resolution differences between displays.


**Figure 1: Task interface, here showing the *complex* task. Participants need to identify and count unique people who sit on the benches positioned below the columns. While this might seem simple, this is a complex spatio-temporal reasoning task as people appear in multiple video windows at very different times. This task requires tracking the entrance and exit of persons across the scene space and time to verify their identity.**

## 3.2 Tasks

To quantitatively assess device effect on video+context panoramas, we need to design tasks for participants to complete. Existing tasks in the literature, e.g., estimating the relative orientation of a boat to infer spatial performance in 2D [34], or using Tri-dimensional chess for 3D [30], are not appropriate for panoramic video imagery which includes both space and time reasoning in hybrid panoramic space. Common actions while exploring video+context panoramas include looking for objects/actions in space and in time, following dynamic events within the place, and identifying when changes happen within specific times or areas. As such, we designed two tasks with different levels of complexity which involve counting and tracking people within several different video windows. To assess performance, we measured the completion time and accuracy expressed as errors in people counts. These metrics are reliable to measure and are not dependent on the display device used.

*Simple Task.* The first task — the *simple task* — asks participants to review 6 videos and count the number of different people who cross between two buildings in a scene (Fig. 2). The videos never fully track a person and do not overlap, and so multiple synchronous videos must be analysed to obtain the correct result. The task may be completed simply and to a high accuracy by manipulating the global timeline only and focusing attention on a specific spatial region in the panorama. Errors are possible by miscounting the number of people who move from building A to building B (e.g., some people exit building A but do not enter building B), with 12 potential errors (manually verified by thorough checking and re-checking). The dataset was collected in a university courtyard. Videos differ in length (2.30–4.00 minutes) but are time sequential, and they cover 125° horizontally within the environment. The task lasts as long as it takes, with an expected time of around 10 minutes.

*Complex Task.* The second task — the *complex task* — required users to browse 20 videos and identify the number of different people who sit on a set of benches within a neo-classical quadrangle (Fig. 1). Videos differ in length (0.20–1.10 minutes), do not overlap in time, and cover the entire horizontal 360° extent of the environment. This

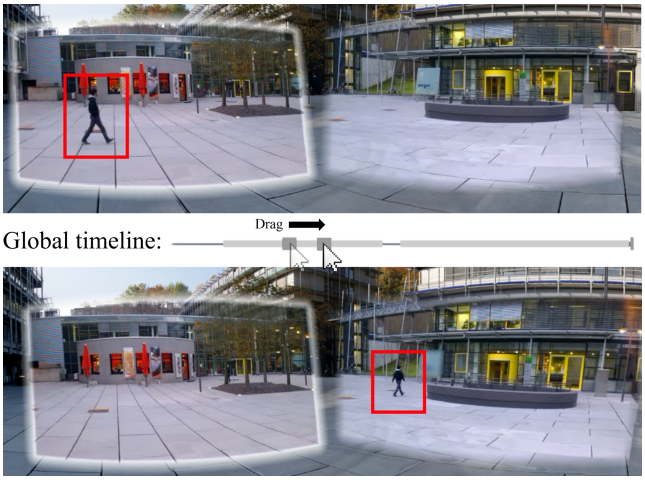


**Figure 2: The *simple* task. Participants must count the number of unique people walking between two buildings. Dragging the global timeline allows tracking people across videos.**

task requires more spatio-temporal reasoning than the simple task as the same people appear in multiple video windows at different times, with some people sitting only near the areas of interest or standing in front of the benches. Participants must focus on parts of the panorama which are farther apart to track the entrances, exits, and identities of people. As some people in the scene leave and return to the benches (mistaken identity counting twice), or as some people pass the benches without sitting down, there are 20 potential erroneous counts (again manually verified). The task lasts as long as it takes, with an expected time of around 10 minutes.

## 3.3 Design Space

With tasks defined, choosing which devices to evaluate from the large number of possible configurations is tricky as each has different properties which might not be directly comparable, and trying to normalize these conditions is difficult. Instead, we choose a systems-level approach, where we try to compare systems which would most likely be used in practice. While this makes it harder to directly compare, instead, it allows us to see the impact of design decisions on participant behaviours with common systems.

Within the design space for this task (Tab. 1), we choose to compare three display scenarios which provide both interesting points in the space and practical systems (Fig. 3):

**Desktop** An exocentric non-immersive display, with mouse control over a cursor.
*Dell U2410 with Belkin Optical Ergo mouse.*

**Tablet** An egocentric non-immersive display, with perspective orientation control through tablet rotation (user togglable), and touchscreen control replacing a cursor. This kind of system is used commonly on mobile devices, e.g., with the Google Maps mobile application, Street View mode can be controlled by device rotation or by touch.
*Acer Iconia W700.*

**HMD** An egocentric immersive display, with perspective orientation control through head rotation and joypad control over a cursor. The HMD is a binocular stereo device; however, we effectively make the display monocular by rendering views of a monocular panorama at infinity.
*Oculus Rift DK1 with Xbox 360 wireless joypad.*

| Properties | Desktop | Tablet | HMD | Spherical | CAVE |
|---|---|---|---|---|---|
| *Input* | | | | | |
| Mouse | ✓ | × | × | × | × |
| Touchscreen | ✓ | ✓ | × | ✓ | × |
| Joypad | ✓ | × | ✓ | ✓ | ✓ |
| Eye-track | ! | ! | ! | ! | ! |
| Hand-track | ! | ! | ! | ! | ! |
| *Display* | | | | | |
| Display size | 24" | 11" | 7" | 16" diam. | 120" |
| Resolution | 1080p | 1080p | ≈500x600 per eye | ≈1024x768 per sphere | 1080p per wall |
| Hor. angle | $\approx 50^\circ$ | $\approx 25^\circ$ | $\approx 100^\circ$ | $\approx 90-135^\circ$ | $\approx 180^\circ$ |
| Immersive | × | × | ✓ | × | ✓ |
| *Modes* | | | | | |
| Egocentric | × | ✓ | ✓ | × | ✓ |
| Exocentric | ✓ | × | × | ✓ | × |
| Perspective | ✓ | ✓ | ✓ | ! | ✓ |
| Equirect. | ✓ | ! | ! | ! | × |
| Flipped space | × | × | × | ! | × |
| In-situ | × | ✓ | × | × | × |

**Table 1: The potential design space of display scenarios. Green marks the chosen display/input combinations, representing combinations likely to be found in practice. Exclamations mark interesting points discussed in the text.**

The desktop and tablet displays represent two commonly used non-immersive solutions, which are baseline for comparison with the immersive HMD device.

Bowman et al. [2] demonstrate that HMDs are a recommended choice when users require strong spatial orientation, outperforming CAVE-like systems. Coupled with their rarity in everyday life, we do not include CAVEs (large projection systems with head-tracking) and instead use a HMD for our egocentric immersive case. We reject tablets physically located in the real world at the same location as the panorama, because there are no other comparison devices. One interesting case is spherical displays, where a world captured from inside-out in a panorama is viewed outside-in looking onto the sphere. This has the effect of flipping spatial relations, where rotation around the sphere reveals imagery in the opposite direction to expectation. Whether this is a problem for spatio-temporal reasoning tasks remains an interesting question; however, we leave this for future work as it is a rare device in practice.

*Inputs.* A change of display often brings with it a required change in input device, making the direct comparison harder. For example, an HMD with physical rotation is difficult to couple with a tethered pointing device, and a handheld tablet makes holding other devices difficult. As we take a systems-level approach, we choose points in the design space where all three display scenarios have different cursor controllers which are the most common input mechanisms for these devices (being mouse, touchscreen, and joypad).

To our knowledge, the literature has no strong conclusions about the absolute effectiveness of these pointing mechanisms, and performance is task and device dependent. There is some evidence to suggest that joypad input has reduced throughput to mouse input (0.69-0.33x bits/s) [33]. There are no wide surveys yet of touch-

**Figure 3: Different display modes used for the study: (*left*) Desktop display with mouse; (*center*) Tablet with rotation and finger orientation controls; (*right*) HMD with head orientation and joypad cursor controls.**

screen and mouse throughputs, but some touchscreen studies have suggested equivalent or faster movement times than with a mouse [29, 6, 10], others that mouse input outperforms touchscreen input when the task requires a single point of contact [15], but also that touchscreens potentially decreased accuracy [27]. In principle, it would be possible to design eye- and hand-tracking systems which are suitable for all of these display types (Tab. 1); however, these technologies are still nascent and uncommon, and currently a consistent integration across displays would be difficult.

As we change input device across displays, we state these important points: First, that interface interaction time is insignificant compared to the expected task completion time. Second, that although our displays have varying angular extent, none have interface elements which fall below the critical angle of difficulty identified by Song et al. [32]. Third, that across our three displays, we ensure that the layout of GUI elements remains consistent by both making the GUI independent of the panoramic view — the GUI moves with your head — and scaled to the display size. To achieve this for the HMD, we render the interface elements onto a plane which follows head rotation at a fixed-depth into the world. For selection, a cursor moves only within this plane, and to mimic a screen, this cursor is bound to the interface elements in much the same way that a mouse is bound to the display's edges. Fourth, that we try to make world rotation amounts consistent across displays to reduce the workload and error rate from the input devices. For both mouse and touchscreen inputs, a display edge-to-edge drag covers 360° degress, and for the HMD there is a 1:1 mapping between head angle and panorama rotation.

*Projections.* For the exocentric desktop case, in systems and the literature, we have seen both equirectangular and perspective projection types commonly used, and no projection is considered optimal. As such, we decided to leave the choice to the user. However, equirectangular projections aren't consistent with egocentric view. For instance, for the HMD, we could present an equirectangular panorama on a plane in a virtual desktop; however, this somewhat defeats the purpose of using an HMD for the immersion and egocentricity. As such, we restrict tablet and HMD devices to use egocentric perspective projections.

# 4. USER STUDY

## 4.1 Procedure

As the number of videos in each task is small ($\leq$ 20) and as human action in video is memorable, there is a large potential for participants to learn the content if we conducted a within-subjects experiment. Instead, we conducted a between-subjects design.

30 participants from the staff and student population at our university performed both tasks using one of the three displays each for a between-subjects design for the display type independent condition, and a within-subjects design for the task. First, each participant was given a detailed description of the interface features, and as much time as they wished to familiarise before the task. Participants could use all features of each system. Then, each task was conducted in series, in a random order, and under no time limit, with a briefing beforehand to explain the task. Following both tasks, the participant completed two questionnaires. While we did not filter the study population for handedness and eyesight, we ensure gender balance was respected. Additionally, the participants were randomly assigned one of the three systems, and there was no mention of the overarching goal of the study.

## 4.2 Hypotheses

Immersive displays such as HMDs might be more suitable to display panoramic representations as they are egocentric, allowing for natural navigation of the environment with head rotation. Tablet devices are less immersive as only a portion of the view is taken up by the virtual window, but tablet use can still be egocentric by rotating the device in space. Desktop displays are exocentric, and so immersion is reduced further. With these premises, and following the results of previous studies which showed that immersion might increase accuracy [30, 21, 34, 20], we hypothesise that accuracy will vary with the level of immersion of the display. We expect the HMD display to be most accurate, the tablet display to be less accurate than the HMD, and for the desktop display to be least accurate.

While input devices differ across displays, we do not expect completion time to vary significantly. As previously discussed, the major workload is in spatio-temporal reasoning and not on interface manipulation. We expect the three conditions to obtain SUS scores relative to their familiarity, with the desktop display obtaining the best score followed, in turn, by the tablet and then the HMD. For the task questionnaire, we expect all three conditions to indicate that the interface was suitable for the task, but we expect exocentric views to be preferred over egocentric views as readability is higher, as per Mulloni et al. [17].

## 4.3 Observations

To complete the tasks, 80% of the population assigned to the desktop condition used equirectangular projection. This finding suggests that participants thought the 360°-at-once projection eased the tasks and, with caution to not infer too much, might suggest that participants felt the equirectangular view conferred more spatio-temporal information than the perspective projection for desktop displays.

| Condition | Simple Task | | | Complex Task | | |
|---|---|---|---|---|---|---|
| | Error | Time (sec.) | Normalised Time | Error | Time (sec.) | Normalised Time |
| Desktop | 0.9 | 373.36 | 0.332 | 1.4 | 469.12 | 0.521 |
| Tablet | 2.1 | 382.50 | 0.340 | 3.6 | 616.00 | 0.684 |
| HMD | 0.8 | 377.40 | 0.336 | 0.9 | 458.80 | 0.509 |

**Table 2: Tasks results. Normalized time is per frame over all video windows. Participants were approximately twice as fast per frame of video at the simple task as it involved constantly comparing two video windows at once within the panorama.**

Regarding tablet users task strategy, 90% of the population preferred to use touch-based rotation rather than orientation rotation navigation. Almost all users first attempted to use the orientation sensor-based navigation, but then switched to touch-based navigation. We discuss the implication of this observation later on.

No one reported eye strain, tiredness, or nausea for the HMD case. HMD users regularly used zoom controls, in contrast to desktop and tablet users who used zoom controls very rarely. This can be explained by the low resolution of the HMD display in comparison with the desktop and tablet displays. Across all conditions, no one struggled to finish the tasks.

# 5. RESULTS

Figure 4 shows box plots for the number of errors committed and the completion time. We computed Analysis of Variance (ANOVA) using SPSS, with the display type used as the single factor, with completion time/counting error as the dependent variables, with post-hoc Tukey tests for pair-wise significance tests ($\alpha < 0.05$), and with no post-hoc pairwise test corrections. No significant difference between desktop and HMD cases existed across all our experiments.

## 5.1 Tasks

Table 2 shows an overview of the results obtained in the two tasks. For the simple task, no significant main effect for device type was found on accuracy ($F(2,1) = 1.581$, $p = 0.224$), with fewer errors for the HMD ($M = 0.8$) and desktop cases ($M = 0.9$) than for the tablet ($M = 2.1$). For completion time, the system used was again not a significant factor ($F(2,1) = 0.13$, $p = 0.987$). The desktop display obtained the lowest mean time ($M = 373$ sec.), followed by the HMD ($M = 377$ sec.) and the tablet ($M = 382$ sec.).

For the complex task, device type was a significant main effect on accuracy ($F(2,1) = 5.173$, $p = 0.013$), with fewer errors for the HMD ($M = 0.9$) and desktop cases ($M = 1.4$) than for the tablet ($M = 3.6$). Post-hoc analysis revealed significant differences between HMD and tablet ($p = 0.049$). For completion time, the system used was a significant factor ($F(2,1) = 3.865$, $p = 0.033$). The HMD case obtained the lowest mean time ($M = 458$ sec.), followed by the desktop ($M = 469$ sec.) and the tablet ($M = 616$ sec.). Post-hoc analysis revealed a significant difference between HMD and tablet ($p = 0.015$), confirming that the HMD allows user to perform their task faster than tablet users. There was also a statistical trend that desktops were more performant than tablets ($p = 0.052$).

## 5.2 Questionnaires

For the system usability scale, both desktop and tablet cases scored above average (SUS = 77.5 and SUS = 76.5 respectively), tailed by the HMD case (SUS = 68). Following the SUS classification technique of Lewis et al. [12] (letter-grade ranks varying from A to F), the desktop and tablet cases are Rank B systems, while the HMD version is a Rank C system. Rank A systems have many promoters, who will definitely use and recommend the product, while rank B systems have a fair number of promoters, who are likely to use and promote the product. All other ranks will only have detractors.

For the task-related questionnaire, across all questions, there were no significant differences (Fig. 5). We conclude that participants felt capable of completing the tasks on all three displays, that all three provided a good sense of orientation, that all three allowed the relative position of videos to be understood, and that all three allowed spatio-temporal reasoning without confusion.

# 6. DISCUSSION

## 6.1 Display Effects

For the simple task, display type was not found to be a significant factor for either completion time or accuracy. We conclude that the task was sufficiently simple that the display type did not make a difference and all three displays are suitable for simple tasks. This shows the importance of using a complex task when assessing spatio-temporal reasoning (supporting [30], contrasting [34]).

For the complex task, the tablet took significantly longer than the HMD and, even though not significant, seems to take longer than the desktop too. Similarly, the tablet is significantly less accurate than the HMD and, even though not significant, seems less accurate than the desktop too. The distributions in both time and accuracy show much larger variance in the tablet case, and this follows the trend in the simple task.

Our hypothesis is not true in our experiment as task accuracy shows that display immersion cannot be considered a significant factor for panoramic imagery. Users were able to achieve equivalent levels of accuracy in both non-immersive (desktop) and immersive (HMD) displays. This suggests that the potentially positive performance effect of immersion in 3D environments does not necessarily extend to panoramas. Further, results from both tasks indicate that egocentric immersive views can be as performant as exocentric non-immersive views in completion time and accuracy.

## 6.2 Tablet

The tablet condition appears worse than the desktop, and was significantly worse than the HMD in the complex task. We suggest that the smaller tablet display, even though it is high resolution, negatively affected spatio-temporal reasoning. From observing participants solving strategies, we did not notice participants zooming or bringing the tablet closer to see more detail. Further, after an initial period of using orientation sensor rotation, nearly all participants switched to touch rotation. This does not explain the longer completion times as, for the simple task, times are comparable across all devices and task order was random. When asking participants to explain why they switched to touch navigation to complete the tasks, participants cited: 1) that camera movement was too tied to device movement, making navigation confusing; 2) that holding the tablet and interacting with the screen was too cumbersome (cf. [39]), and 3) that the device was too heavy to hold in this way for long periods of time.

One might think that tablet resolution was a factor. For 1080p at 25 degrees field of view, each pixel on the tablet equals 0.6 arcminutes of view, in contrast with human eye acuity of approximately 1.2 arcminutes. As the tablet is mobile, this extra resolution could be viewed by simply moving the tablet closer. However, in general,

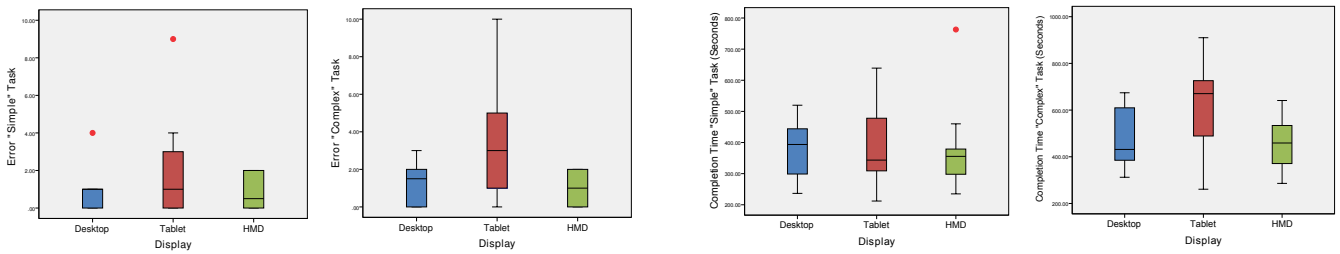


**Figure 4: Box plots for counting errors (*left*) and completion time (*right*) for each display type and task. Red dots are outliers.**

| Task-related question | Desktop | Tablet | HMD |
|---|---|---|---|
| Q1: Easy to complete tasks | 4.0 | 4.5 | 3.66 |
| Q2: Understood video orientation in space | 4.7 | 3.9 | 3.8 |
| Q3: Understood relative video position | 4.4 | 4.2 | 3.6 |
| Q4: Understood space-time video overlap | 4.3 | 4.1 | 4.0 |
| Q5: Understood temporal order of videos | 3.4 | 3.3 | 3.3 |
| Q6: Environment representation confused | 3.3 | 3.1 | 2.5 |
| Q7: System has enough functions for tasks | 4.4 | 4.1 | 4.0 |
| Q8: #videos made remembering things hard | 2.4 | 1.6 | 1.9 |
| Overall mean | 3.86 | 3.6 | 3.34 |

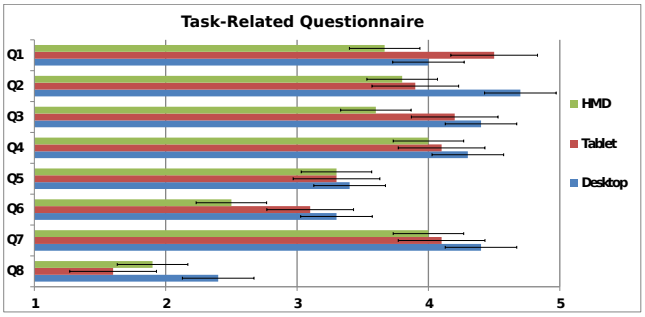


**Figure 5: Mean and variance plot for the task-related questionnaire. We assume Likert ordinal data was fairly interpreted as an interval scale, with text labels ranging from *strongly disagree* to *strongly agree*. The scale for negative questions was reversed for mean computation.**

this is a moot point and does not hinder performance, because the focus areas of the task — the people in scenes — with no zoom, are typically 10-30 pixels wide, and 250–2250 pixels in area.

Interestingly, the task questionnaire suggests participants felt the tablet was just as capable as the desktop, and the SUS scores suggest participants felt it was just as usable, too. This does not align with real task performance, which was reduced for the complex task. We suggest this is a familiarization issue, as participants were comfortable in general with tablets. This 'false sense of security' is potentially dangerous if tablets were to be used for critical panoramic review tasks (see Section 6.4).

## 6.3 HMD

The HMD performed similarly to the desktop, and significantly better than the tablet, However, the questionnaires scores suggest that users found it less capable, and the SUS scores suggest that users found it less usable. While one might think that the HMD was rated as less capable or usable for human reasons (eye strain, tiredness, nausea, general uncomfort), our participants reported no such issues. Instead, we suggest that this is a familiarization issue again, but now the reverse effect where the novelty of the device induces caution in qualitative assessment. However, given the equivalent performance to the desktop case, there is no reason to suggest that the HMD interface is a compromise for our tasks. Again, one might think that resolution would be an issue, as the HMD has 10x lower perceived resolution (with 12 arcminutes, in contrast to the desktop with 1.2 arcminutes). However, with simple zoom controls, the two display types performed similarly. This suggests that the tasks performance does not differ simply from a change in resolution perception.

The lack of benefit from using an HMD over a desktop in our experiment, expected from the immersion suggestions from virtual environment works [30], is unlikely to be attributed to the difference between rendered and photographed views, as Willemsen and Gooch suggest [40]. Instead, we suggest this parity comes from the added warped field of view provided by the exocentric equirectangular projection on the desktop.

## 6.4 Applications and Design Implications

We wish to discuss the potential generalization of our results. Many works in this field (and others) provide evidence for more general conclusions from a single experiment [30, 34], which helps form a body of evidence within the literature for the general conclusion. We have seen effects in specific tasks that are limited to panoramic video imagery; however, for corroboration, the existing work concerning devices and panoramic imagery is limited. As explained in the introduction, this is imagery used daily by thousands of people, and so we describe applications and provides design implications of our results for these applications. However, we caution the reader from drawing implications beyond our scope, and we anticipate a continued scientific discussion on the effects of panoramic imagery.

Surveillance is an important application which commonly produces data from cameras mounted to pan and tilt heads, and this exactly fits our scenario of video+context. Critical tasks might include reviewing videos over time and space for suspicious behavior, or reviewing videos over time and space to identify and localize a person of interest. Our findings are directly applicable here.

We imagine the virtual tourism industry will implement new systems to display videos of the time and space of a place, and our implications inform exploration interfaces for these applications. For instance, our experimental setup well-models a system where users upload their own personal videos of a famous place, to be explored as part of an online collection of all videos uploaded of that place within a context — say, an enhanced, user-driven Google Street View. This might be extended to include games, similar to existing panoramic games such as GeoGuessr or Myst 3, such as video treasure hunts or puzzles.

Our final application is panoramic telepresence applications [18, 22]. Tablets are common communication devices, but for panoramic representations our experiment provides evidence that desktops and HMDs are better systems. Telepresence can become a critical application just by who is involved — doctors, law enforcement, military — for direct communication and collaboration, review, and remote control uses.

From our experience, task-based study, and questionnaires, we suggest implications of our study for these and other applications with similar video+context components:

- Participants preferred exocentric equirectangular projections over perspective projections on desktops (as per Mulloni et al. [17]). This projection type seems to be an appropriate default for desktop systems.
- Participants preferred touch rotation controls over arm-based orientation controls for tablets, as it is difficult to both orient and manipulate on-screen elements. The ability to pause orientation control is necessary. Even with this option, for our reasoning tasks, participants reverted permanently to touch rotation. In-situ browsing and augmented reality situations may provoke a different response; however, in general, arm-based orientation controls are not recommended for tasks requiring long periods of concentration as they are tiring.
- For our more complex panoramic spatio-temporal reasoning task, tablets are less suitable than either desktops or HMDs. Furthermore, users thought the tablet was as usable and as capable as the desktop, suggesting users were not aware of a performance deficit. Critical panoramic review applications should be cautious about using tablets without further investigation.
- HMDs appear to be a viable alternative to desktops for our panoramic spatio-temporal reasoning tasks. While, to participants, they appear to be less usable and less capable, this seems not to be the case, and implementing applications should be aware of this potential negative bias. We expect that, after a longer period of familiarization, this effect will diminish.
- It is important to provide zoom controls to overcome the comparatively low-resolution of some HMDs.
- Comparing all three displays for complex tasks, the higher FOV displays (desktop, HMD) were more successful (supporting [34]).

# 7. CONCLUSION

We have investigated the effect of different devices on panoramic video+context task performance. To create one simple and one complex reasoning task, we exploited new panoramic context and video window ideas and created an adaptive multi-display interface. We conducted a user study with desktop, tablet, and HMD devices covering exocentric and egocentric modes. We discovered that desktop and HMD devices perform comparably, even if users feel the HMD is less capable and less usable. We find that tablet displays are significantly less performant in our complex spatio-temporal reasoning task, even though participants found them as capable and usable as a desktop. These results form implications for panoramic imagery interfaces and applications.

Finally, we do not consider binocular vision in this study. Performance improvements from immersive stereo displays have been shown for 3D environments, and binocular depth cues are one factor in immersion that we did not replicate here. Stereo panoramic imagery exists, though it is uncommon, and future work should investigate the effect of this cue.

# 8. REFERENCES


[1] S. Benford, C. Brown, G. Reynard, and C. Greenhalgh. Shared spaces: Transportation, artificiality, and spatiality. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, CSCW '96, pages 77–86, New York, NY, USA, 1996. ACM.

[2] D. A. Bowman, A. Datey, Y. S. Ryu, U. Farooq, and O. Vasnaik. Empirical comparison of human behavior and performance with different display devices for virtual environments. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, pages 2134–2138, 2002.

[3] A. Braun and R. McCall. Short paper: user study for mobile mixed reality devices. In *Proceedings of the 16th Eurographics conference on Joint Virtual Reality of EGVE - EuroVR - VEC*, EGVE - JVRC'10, pages 89–92, Aire-la-Ville, Switzerland, Switzerland, 2010. Eurographics Association.

[4] S. E. Chen. Quicktime VR: An image-based approach to virtual environment navigation. In *Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, pages 29–38, New York, NY, USA, 1995. ACM.

[5] A. Dalvandi, B. E. Riecke, and T. Calvert. Panoramic video techniques for improving presence in virtual environments. In *Proceedings of the 17th Eurographics conference on Virtual Environments; Third Joint Virtual Reality*, EGVE - JVRC'11, pages 103–110, Aire-la-Ville, Switzerland, Switzerland, 2011. Eurographics Association.

[6] C. Forlines, D. Wigdor, C. Shen, and R. Balakrishnan. Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 647–656, New York, NY, USA, 2007. ACM.

[7] L. Gamberini, A. Spagnolli, L. Prontu, S. Furlan, F. Martino, B. R. Solaz, M. Alcañiz, and J. A. Lozano. How natural is a natural interface? an evaluation procedure based on action breakdowns. *Personal Ubiquitous Comput.*, 17(1):69–79, Jan. 2013.

[8] J. F. Jensen. From "Flatland" to "Spaceland" spatial representation, enunciation and interaction in 3D-virtual worlds. *WebNet Journal: Internet Technologies, Applications & Issues*, 1(1):44–57, 1999.

[9] J. F. Jensen, editor. *Virtual Space: Spatiality in Virtual Inhabited 3D Worlds*. Springer-Verlag, London, UK, UK, 2002.

[10] N. Jochems, S. Vetter, and C. Schlick. A comparative study of information input devices for aging computer users. *Behaviour & Information Technology*, 32(9):902–919, 2013.

[11] J. F. Lapointe, P. Savard, and N. G. Vinson. A comparative study of four input devices for desktop virtual walkthroughs. *Comput. Hum. Behav.*, 27(6):2186–2191, nov 2011.

[12] J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009*, HCD 09, pages 94–103, Berlin, Heidelberg, 2009. Springer-Verlag.

[13] A. Lippman. Movie-maps: An application of the optical videodisc to computer graphics. *SIGGRAPH Comput. Graph.*, 14(3):32–42, July 1980.

[14] A. Majumder, W. B. Seales, M. Gopi, and H. Fuchs. Immersive teleconferencing: a new algorithm to generate seamless panoramic video imagery. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 169–178. ACM, 1999.

[15] S. Meyer, O. Cohen, and E. Nilsen. Device comparisons for goal-directed drawing tasks. In *Conference Companion on Human Factors in Computing Systems*, CHI '94, pages 251–252, New York, NY, USA, 1994. ACM.

[16] D. Mizell, S. Jones, M. Slater, and B. Spanlang. Comparing immersive virtual reality with other display modes for visualizing complex 3D geometry. Technical report, University College London, 2002.

[17] A. Mulloni, H. Seichter, A. Dünser, P. Baudisch, and D. Schmalstieg. Panoramic overviews for location-based services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2565–2568, New York, NY, USA, 2012. ACM.

[18] J. Norris, H. Schnädelbach, and G. Qiu. Camblend: an object focused collaboration tool. In *Proceedings of SIGCHI '12*, New York, NY, USA, 2012. ACM.

[19] C. S. Ozbek, B. Giesler, and R. Dillmann. Jedi training: playful evaluation of head-mounted augmented reality display systems. *Proc. SPIE*, 5291:454–463, 2004.

[20] E. Patrick, D. Cosgrove, A. Slavkovic, J. A. Rode, T. Verratti, and G. Chiselko. Using a large projection screen as an alternative to head-mounted displays for virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '00, pages 478–485. ACM, 2000.

[21] R. Pausch, M. Shackelford, and D. Proffitt. A user study comparing head-mounted and stationary displays. In *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, pages 41–45, 1993.

[22] F. Pece, W. Steptoe, F. Wanner, S. Julier, T. Weyrich, J. Kautz, and A. Steed. Panoinserts: Practical spatial teleconferencing. In *Proceedings of SIGCHI '13*, 2013.

[23] S. Pirk, M. F. Cohen, O. Deussen, M. Uyttendaele, and J. Kopf. Video enhanced gigapixel panoramas. *SIGGRAPH Asia 2012 Technical Briefs on*, 2012.

[24] N. F. Polys, S. Kim, and D. A. Bowman. Effects of information layout, screen size, and field of view on user performance. In *in Information-Rich Virtual Environments, Proceedings of ACM Symposium on Virtual Reality Software and Technology*, pages 46–55, 2005.

[25] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert. The perception of egocentric distances in virtual environments - a review. *ACM Comput. Surv.*, 46(2):23:1–23:40, Dec. 2013.

[26] B. E. Riecke, J. Schulte-Pelkum, and H. H. Buelthoff. Perceiving simulated ego-motions in virtual reality: comparing large screen displays with hmds. *Proc. SPIE*, 5666:344–355, 2005.

[27] F. Sasangohar, I. MacKenzie, and S. Scott. Evaluation of mouse and touch input for a tabletop display using Fitts' reciprocal tapping task. In *Proceedings of the 53rd Annual Meeting of the Human Factors and Ergonomics Society (HFES 2009)*, volume 2, pages 839–843, 2009.

[28] R. Schroeder, A. Steed, A.-S. Axelsson, I. Heldal, A. Abelin, J. Widestrom, A. Nilsson, and M. Slater. Collaborating in networked immersive spaces: as good as being there together? *Computers and Graphics*, 25(5):781 – 788, 2001.

[29] A. Sears and B. Shneiderman. High precision touchscreens: design strategies and comparisons with a mouse. *Int. J. Man-Mach. Stud.*, 34(4):593–613, Apr. 1991.

[30] M. Slater, V. Linakis, M. Usoh, R. Kooper, and G. Street. Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. In *ACM Virtual Reality Software and Technology (VRST)*, pages 163–172, 1996.

[31] M. Slater and M. Usoh. Presence in immersive virtual environments. In *Virtual Reality Annual International Symposium, 1993., 1993 IEEE*, pages 90–96, 1993.

[32] H. Song. Updating Fitts' law to account for restricted display field of view conditions. *International Journal of Human-Computer Interaction*, 28(4):269–279, 2012.

[33] R. W. Soukoreff and I. S. MacKenzie. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in {HCI}. *International Journal of Human-Computer Studies*, 61(6):751 – 789, 2004.

[34] D. S. Tan, D. Gergle, P. Scupelli, and R. Pausch. With similar visual angles, larger displays improve spatial performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 217–224. ACM, 2003.

[35] W.-K. Tang, T.-T. Wong, and P.-A. Heng. The immersive cockpit. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pages 658–659, New York, NY, USA, 2002. ACM.

[36] M. Ten Koppel, G. Bailly, J. Müller, and R. Walter. Chained displays: configurations of public displays can be used to influence actor-, audience-, and passer-by behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 317–326, New York, NY, USA, 2012. ACM.

[37] J. Tompkin, F. Pece, R. Shah, S. Izadi, J. Kautz, and C. Theobalt. Video collections in panoramic contexts. In *UIST*, 2013.

[38] M. Tsang, G. W. Fitzmaurice, G. Kurtenbach, A. Khan, and B. Buxton. Boom chameleon: Simultaneous capture of 3D viewpoint, voice and gesture annotations on a spatially-aware display. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*, UIST '02, pages 111–120, New York, NY, USA, 2002. ACM.

[39] J. Wagner, S. Huot, and W. Mackay. BiTouch and BiPad: Designing Bimanual Interaction for Hand-held Tablets. In *CHI'12 - 30th International Conference on Human Factors in Computing Systems - 2012*. ACM SIGCHI, ACM Press, 2012.

[40] P. Willemsen and A. A. Gooch. Perceived egocentric distances in real, image-based, and traditional virtual environments. In *Proceedings of the IEEE Virtual Reality Conference 2002*, VR '02, pages 275–. IEEE Computer Society, 2002.

[41] J. Wither, S. DiVerdi, and T. Hollerer. Evaluating display types for ar selection and annotation. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 95–98, 2007.

[42] B. Witmer and P. Kline. Judging perceived and traversed distance in virtual environments. *Presence*, 7(2):144–167, April 1998.